GeomPrompt: Geometric Prompt Learning for RGB-D Semantic Segmentation Under Missing and Degraded Depth

Krishna Jaganathan, Patricio Vela
Georgia Institute of Technology
{kjaganathan7, pvela}@gatech.edu

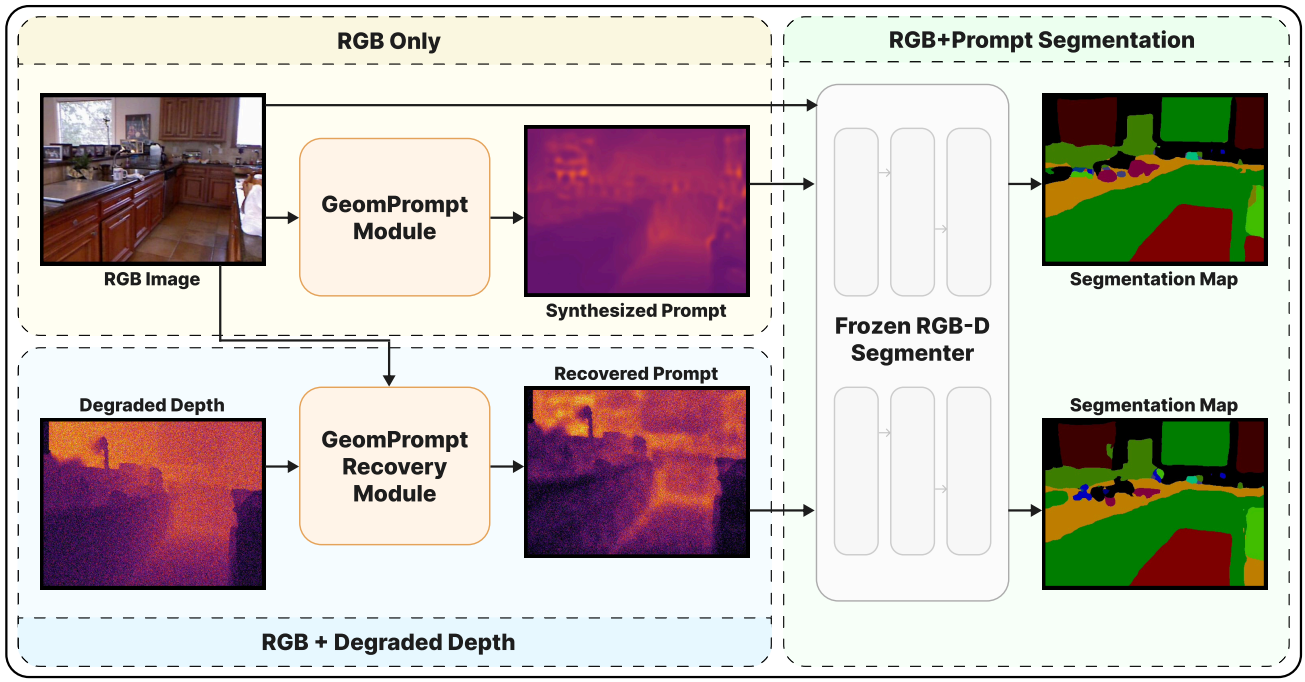


Figure 1. *GeomPrompt* synthesizes a geometric prompt from an RGB image for downstream semantic segmentation on a frozen RGB-D segmenter in the case of missing depth. *GeomPrompt-Recovery* extends this by allowing for corrections on existing degraded depth inputs.

Abstract

Multimodal perception systems for robotics and embodied AI often assume reliable RGB-D sensing, but in practice, depth is frequently missing, noisy, or corrupted. We thus present GeomPrompt, a lightweight cross-modal adaptation module that synthesizes a task-driven geometric prompt from RGB alone for the fourth channel of a frozen RGB-D semantic segmentation model, without depth supervision. We further introduce GeomPrompt-Recovery, an adaptation module that compensates for degraded depth by predicting the fourth channel correction relevant for the frozen segmenter. Both modules are trained solely with downstream segmentation supervision, enabling recovery of the geometric prior useful for segmentation, rather than estimating depth signals. On SUN RGB-D, GeomPrompt improves over RGB-only inference by +6.1 mIoU on DFormer and +3.0 mIoU on GeminiFusion, while remaining competitive with strong monocular depth estimators. For degraded depth, GeomPrompt-Recovery consistently improves robustness, yielding gains up to +3.6 mIoU under severe depth corruptions. GeomPrompt is also substantially more efficient than monocular depth baselines, reaching 7.8 ms latency versus 38.3 ms and 71.9 ms. These results suggest that task-driven geometric prompting is an efficient mechanism for cross-modal compensation under missing and degraded depth inputs in RGB-D perception.

Supported in part by NSF Award #2345057.
Project page: https://geomprompt.github.io

1. Introduction

Semantic segmentation is a core perception primitive for embodied systems because it converts raw sensor observations into structured scene understanding that supports action. In robotic settings, segmentation provides object-level and layout-level priors that help agents reason about what is present, where it is, and how to move or interact safely in cluttered environments. These semantic cues are useful for downstream tasks such as object-goal navigation, semantic mapping, and mobile manipulation, where success depends not only on recognizing categories but also on grounding those categories in spatially meaningful scene structure [11, 43, 51].

A common way to strengthen semantic segmentation is to incorporate geometry, especially through RGB-D sensing. Depth provides complementary structural information that can improve pixel-level prediction beyond what is available from appearance alone, and RGB-D segmentation has thus become a standard bimodal setting in both the segmentation and robotics literature [27, 55, 65, 66]. In this regime, depth is valuable as a prior that can sharpen boundaries, separate objects with similar appearance, and improve scene layout reasoning [55, 66].

However, this reliance on depth creates a practical mismatch at deployment time. In real robotic systems, depth can be unavailable, spatially incomplete, noisy, quantized, or otherwise unreliable due to sensor failures, difficult materials, environmental conditions, or range and hardware limitations [22, 23, 33, 50]. Reflective and transparent surfaces, for example, are well known to produce corrupted or missing measurements, and even when depth is present, its quality can vary substantially across sensors and operating conditions [23, 26, 50]. As a result, the settings in which geometric cues are most helpful for perception are often exactly those in which depth is hardest to trust [33].

Existing alternatives partially resolve this gap. Some approaches use depth or geometry as privileged information during training, while others insert monocular depth estimation as an intermediate step before segmentation [19, 20, 25, 28, 64]. These strategies can be effective, but they typically still depend on explicit geometric supervision, depth-oriented pretraining, synthetic RGB-D data, pseudo-depth pipelines, or additional estimation modules whose objective is to reconstruct metric or proxy geometry rather than directly optimize the downstream segmentation task [25, 29, 64]. This leaves open a simpler question: When a strong RGB-D segmenter expects geometric input, can we learn to supply only a geometry-like signal sufficient for the segmenter, without supervising that signal as depth at all [20, 28]?

We address this question with *GeomPrompt*, a lightweight module that learns to synthesize a geometric prompt from RGB for a frozen RGB-D segmenter. Rather than reconstructing metric depth, GeomPrompt learns the geometry-like signal that is useful for the downstream multimodal model, through segmentation-only supervision. We further introduce *GeomPrompt-Recovery*, a conditioned variant for the complementary setting in which depth is present but unreliable. Given RGB together with degraded depth, it predicts a bounded residual correction that restores task-relevant geometric information for the frozen segmenter under segmentation-only supervision. In this view, GeomPrompt and GeomPrompt-Recovery act as lightweight mechanisms for cross-modal guidance and compensation in RGB-D perception under sensor failure.

This perspective reframes the problem from estimating depth to recovering the geometric prior relevant for RGB-D segmentation under sensor failure. Empirically, our results show that this simple prompting view yields a resource-aware alternative to explicit monocular depth estimation, as it preserves frozen RGB-D backbones, improves performance under missing and degraded depth, and remains lightweight enough to be attractive for resource-constrained embodied perception.

Our core contributions are as follows.

1. *GeomPrompt*, a lightweight cross-modal adaptation module for frozen RGB-D segmentation under missing depth, which synthesizes a task-relevant geometric prompt from RGB alone.
2. *GeomPrompt-Recovery*, a degradation-aware recovery module that predicts task-relevant corrections for degraded depth rather than reconstructing metric depth.
3. Evidence that segmentation-only supervision can support robust multimodal segmentation under sensor unreliability, improving missing and degraded modality performance while remaining efficient.

2. Related Work

2.1. RGB-D Semantic Segmentation

Recent RGB-D semantic segmentation methods improve performance by fusing depth with RGB through dual-stream encoders, cross-modal attention, and geometry-aware priors. Representative approaches such as DFormer [65], GeminiFusion [27], CMX [67], DFormerv2 [66], and AsymFormer [14] span efficient multimodal fusion, explicit geometric priors, and real-time asymmetric designs, and together establish strong baselines for RGB-X segmentation.

A broader family of works further explores scale-invariant depth encoding, attention-based fusion, and efficient backbone design across RGB-D architectures [7, 35, 47, 54, 56, 57, 59, 60, 62, 72]. Despite their strong empirical performance, these methods are typically designed under the assumption that depth is available and sufficiently reliable at inference time.

2.2. Geometry as Privileged Information

To reduce reliance on test-time depth, several approaches use geometry as privileged information available only during training. Methods such as Geometry-Aware Distillation [28] and Hard Pixel Mining for Depth Privileged Semantic Segmentation [20] use distillation or depth privileged supervision to improve RGB segmentation without requiring depth at inference.

Related works also leverage depth or synthetic geometric structure for domain adaptation and unsupervised representation learning [8, 32, 48, 53]. In contrast, GeomPrompt does not predict, regress, reconstruct, or distill depth explicitly. Instead, it learns a task-relevant auxiliary prompt for a frozen RGB-D segmenter using segmentation loss alone.

2.3. Estimated Depth as Proxy Geometry

Another common strategy is to synthesize or estimate proxy depth to compensate for missing or unreliable sensor measurements. This includes modern supervised and metric monocular depth estimators such as Depth Anything V2 [64], Metric3Dv2 [25], and UniDepthV2 [39], as well as self-supervised approaches such as Monodepth2 [19]. Many segmentation pipelines follow a two-stage paradigm in which depth is first estimated and then consumed by the segmentation model [6, 21], while others use depth-oriented pretraining before segmentation fine-tuning [31].

Additional methods introduce pseudo-depth or estimated depth assistance through diffusion models or semi-supervised auxiliary training [24, 30, 52, 63]. By contrast, GeomPrompt does not rely on explicit depth prediction, depth-oriented pretraining, real or synthetic depth supervision, or photometric self-supervision.

2.4. Missing Modality Robustness

A growing body of literature studies semantic segmentation under missing modalities or degraded sensor inputs. Techniques such as Depth Removal Distillation [16] explicitly reduce dependence on depth, while methods such as M3L [36], MAGIC [71], and Delivering Arbitrary-Modal Semantic Segmentation [68] train multimodal architectures to operate under missing or varying modality availability. Recent benchmarks further highlight the fragility of multimodal segmentation systems under realistic sensor failures and modality corruption [33].

Related robustness architectures address fusion robust to noise, along with degraded depth handling and diffusion-based refinement within RGB-D segmentation pipelines [4, 7, 59, 72]. In contrast to methods that train or adapt the multimodal backbone itself for robustness, both GeomPrompt and GeomPrompt-Recovery act as lightweight adaptation modules for a frozen RGB-D segmenter under missing (GeomPrompt) or corrupted/quantized (GeomPrompt-Recovery) depth.

3. Methodology

3.1. Prompting Framework

We consider a frozen RGB-D segmenter S deployed under two conditions, namely missing depth, where only RGB x is available, and degraded depth, where RGB is paired with corrupted depth $\tilde{d}$. Rather than reconstructing metric depth, we learn a task-driven geometric prompt p^* in the depth input space expected by S. In the missing depth setting, GeomPrompt predicts $p^* = G(x)$, and in the degraded depth setting, GeomPrompt-Recovery predicts $p^* = G_R(x, \tilde{d})$. The final segmentation prediction is given by $\hat{y} = S(x, p^*)$. Both modules are trained only with segmentation supervision from y, without depth supervision.

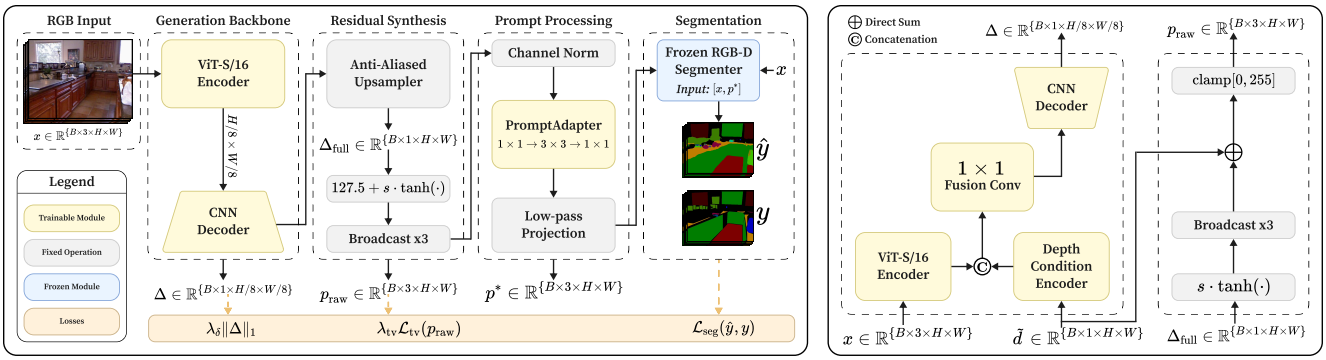
3.2. GeomPrompt Architecture

GeomPrompt accepts an ImageNet-normalized RGB image [12] $x \in \mathbb{R}^{3\times H\times W}$ as input. The primary output of the module is the normalized prompt p^*. Auxiliary outputs generated during the forward pass include a low-resolution residual map Δ and the raw continuous prompt p_{raw}.

To extract rich visual semantics, the input image is processed by a ViT-S/16 encoder [13]. Following the standard vision transformer forward pass, prefix tokens (such as the class token) are discarded [42]. The remaining patch tokens are then reshaped back into a 2D spatial feature grid, preserving the spatial layout of the original image for dense prediction [42].

The spatial feature grid is passed to a lightweight CNN decoder consisting of an initial $\times 2$ bilinear upsampling stage followed by two 3×3 Conv-BN-ReLU blocks and a final 1×1 projection to a single-channel residual map. This yields a compact decoder that maps reassembled ViT features to a low-resolution geometric residual [42]. Rather than predicting a full-resolution depth map directly, the decoder predicts a low-resolution residual Δ at a spatial scale of $H/8 \times W/8$. This residual is then progressively upsampled to the full $H \times W$ resolution using a fixed anti-aliased upsampler, yielding Δ_{full}. By combining a low-resolution decoder with anti-aliased upsampling, the network suppresses spectral and checkerboard artifacts and favors more stable and smoother geometric structures over noisy high frequency per-pixel details [2, 38, 41].

To ensure the synthesized prompt remains stable and conforms to the expected input space of the frozen segmenter, the upsampled residual undergoes a strict parameterization process. The residual is bounded using a scaled hyperbolic tangent function, $s\tanh(\cdot)$, and centered around a neutral gray prior of 127.5. The resulting single-channel map is expanded to three channels, normalized, refined by a residual PromptAdapter, and finally passed through a hard low-pass projection before entering the segmenter. The PromptAdapter is a lightweight convolutional resid-



(a) Overview of the GeomPrompt architecture.

(b) Depth handling in GeomPrompt-Recovery.

Figure 2. Model architecture diagrams. (a) Overview of the GeomPrompt architecture. (b) Depth handling in GeomPrompt-Recovery.

ual module operating in 3-channel prompt space, with a $1 \times 1 \quad 3 \rightarrow 16$ projection, a $3 \times 3 \quad 16 \rightarrow 16$ convolution, and a final $1 \times 1 \quad 16 \rightarrow 3$ projection. Its last layer is zero-initialized so that the module starts as an identity mapping, allowing it to learn only segmenter-specific corrections to the normalized prompt. The low-pass projection applies average-pooling downsampling followed by bilinear upsampling with a factor of 2. This suppresses high-frequency prompt artifacts and enforces a smoother geometric signal [70]. This parameterization can be formalized as:

$$\begin{aligned} \Delta_{\text{full}} &= \mathcal{U}(\Delta) \\ p_{\text{raw}} &= 127.5 + s\tanh(\Delta_{\text{full}}) \\ p^* &= \Pi(\mathcal{A}(\mathcal{N}(p_{\text{raw}}))) \end{aligned} \tag{1}$$

Where $\mathcal{U}$ denotes the fixed anti-aliased upsampler, $\mathcal{N}$ represents the normalization, $\mathcal{A}$ is the PromptAdapter, and Π is the hard low-pass projection. We replicate the prompt to three channels before applying these operations, yielding a uniform interface for all frozen segmenters. Segmenters that consume three channel geometric input use the prompt directly, while those expecting one channel apply their standard segmenter-specific channel reduction during preprocessing. Figure 2a details the end-to-end architecture of GeomPrompt.

3.3. GeomPrompt-Recovery Architecture

The architecture of GeomPrompt-Recovery is structured as a dual-pathway system that fuses multimodal features before predicting a structural correction. It retains the same primary RGB ViT branch utilized in the base GeomPrompt module. To process the degraded depth, we introduce an additional lightweight depth-condition encoder, where the degraded depth is first replicated to 3 channels and passed through a lightweight 4-layer 3×3 stride-2 CNN ($3 \rightarrow 32 \rightarrow 64 \rightarrow 64 \rightarrow 64$), producing a local structural feature map aligned to the $H/16 \times W/16$ token grid of the RGB ViT branch. The extracted RGB and depth features are concatenated along the channel dimension and subsequently fused using a 1×1 convolutional layer. This fused representation is then passed through the identical decoder and anti-aliased upsampler stack used in GeomPrompt to predict a residual correction map.

GeomPrompt-Recovery predicts a bounded residual correction that is applied directly to the incoming corrupted depth, $\tilde{d}$, in its raw continuous space. This additive correction process is formulated as:

$$\begin{aligned} \text{corr} &= s\tanh(\Delta_{\text{full}}) \\ p_{\text{raw}} &= \text{clamp}(\tilde{d} + \text{corr}, 0, 255) \end{aligned} \tag{2}$$

Where Δ_{full} is the progressively upsampled output from the decoder and s dictates the bounds of the scaled hyperbolic tangent function. Following this residual correction and clamping, the recovered prompt p_{raw} undergoes the same downstream normalization, PromptAdapter refinement, and hard low-pass projection as defined in the base GeomPrompt module to produce p^*.

For principled and stable optimization, the correction head of the decoder is zero-initialized, so the model begins its training regime as an approximate identity mapping with respect to the supplied broken depth. The network is thus forced to learn purposeful deviations from the corrupted depth only when those deviations are supported by the downstream segmentation loss.

GeomPrompt-Recovery thus acts as a corruption-aware repair mechanism, learning to correct noisy or missing depth information only insofar as that recovery directly improves the semantic parsing capabilities of the frozen RGB-D segmenter. Figure 2b details the degraded depth handling that is present in GeomPrompt-Recovery.

3.4. Training Objective and Protocol

Across both variants of our method, the downstream RGB-D segmentation backbone remains entirely frozen; only the parameters of the prompt generation module are optimized. The primary driving signal is the segmentation

loss on the final semantic prediction. To encourage stable and well-behaved prompt generation, we augment this primary objective with two regularizers: a Total Variation (TV) smoothness penalty applied to the raw continuous prompt, and an L1 magnitude penalty applied to the low-resolution residual. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{seg}}(\hat{y}, y) + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}(p_{\text{raw}}) + \lambda_{\delta}\|\Delta\|_1 \tag{3}$$

Here, $\mathcal{L}_{\text{seg}}$ represents the Online Hard Example Mining (OHEM) Cross-Entropy loss, and λ_{tv} and λ_{δ} are empirical weighting coefficients.

To maintain stability during training, the residual scaling factor s, which bounds the prompt generation, is progressively ramped over training epochs. This allows the model to smoothly transition from predicting a safe, uniform gray prior to generating structurally complex, learned geometric features.

To equip the GeomPrompt-Recovery module with deployment robustness, we train it using a synthesized distribution of degraded depth inputs. We stochastically apply a diverse suite of spatial and systemic corruptions to GT depth, with severities dynamically varied across training iterations to expose the network to a broad spectrum of potential sensor failures.

4. Experiments

4.1. Implementation Details

We train GeomPrompt and GeomPrompt-Recovery for 300 epochs using AdamW with a poly learning-rate schedule (power 0.9), 10 warmup epochs, weight decay 0.01, and separate learning rates of $3 \cdot 10^{-5}$ for the encoder and $1 \cdot 10^{-4}$ for the decoder. Training uses an effective batch size of 32 on $8\times$ A40 GPUs. We set $\lambda_{tv} = 10^{-5}$ and $\lambda_{\delta} = 5 \times 10^{-4}$, and linearly ramp the residual bound s from 15 to 80 over all epochs. All training images are processed with random resize and random crop to 480×480. For GeomPrompt-Recovery, each training image is either kept clean with probability 0.2 or assigned exactly one uniformly sampled corruption from {*quantize, hole, dropout, noise, blur, banding, scale shift*}. Corruption severity is sampled uniformly from $[0.10, 0.90]$. GeomPrompt outputs $p_{\text{raw}} \in [0, 255]$, which we normalize using the DFormer [65] depth statistics, namely $\mu = [0.48, 0.48, 0.48]$ and $\sigma = [0.28, 0.28, 0.28]$ as $\mathcal{N}(p_{\text{raw}})$ for consistency across backbones.

4.2. Experimental Setup

We evaluate our approach on the standard test split of the SUN RGB-D dataset [49]. We employ DFormer [65] and GeminiFusion [27] as our base pretrained RGB-D segmentation models. These two architectures serve as complementary testbeds because they differ both in how they incorporate geometry and in how they are evaluated, as DFormer uses a dedicated hierarchical RGB-D encoder with four-stage multi-scale features and geometry-aware RGB-D blocks that combine global-awareness and local-enhancement attention, whereas GeminiFusion uses lightweight pixel-wise fusion over aligned cross-modal features within a shared four-stage transformer backbone [27, 65].

Additionally, DFormer is evaluated with multi-scale flip inference, while GeminiFusion's evaluation is reported without multi-scale or flip test-time augmentation, allowing us to test GeomPrompt under both architectural and evaluation protocol variation [27, 65].

To rigorously evaluate GeomPrompt, we compare its synthesized prompts against ground truth depth maps from SUN RGB-D as an upper bound, along with monocular depth estimators and an RGB-only baseline.

For our depth estimators, we use state-of-the-art zero-shot monocular depth models, specifically Metric3Dv2 (ViT-small) [25] and the standard and Hypersim [44] checkpoints of Depth Anything 2 (DA2) (ViT-base) [64]. We evaluate on both these checkpoints as the Hypersim checkpoint yields metric depth whereas the standard checkpoint yields relative depth. For our RGB-only baseline, we set all depth pixels to zero. This serves as a control for a complete lack of geometric signal.

Our primary evaluation metric is mean Intersection over Union (mIoU), supported by pixel accuracy (PA) as a secondary metric. To ensure a fair comparison, we adhere to a strict shared evaluation principle, as for a given model family, the RGB preprocessing, model backbone, and inference configuration, such as test-time augmentation and scaling, remain identical. The only variable altered during evaluation is the source of the depth channel input.

4.3. RGB Geometric Prompting

For the evaluation on DFormer, inference is conducted using standard multi-scale and flip test-time augmentation at scales of $\{0.5, 0.75, 1.0, 1.25, 1.5\}$, as is done in the original paper [65]. We apply this on the whole image at original size, which reproduces the reported mIoU for GT depth. We use the DFormer-Base variant of DFormer. The GeminiFusion evaluation protocol uses a single-scale, no-flip inference pipeline without test-time augmentation, with images resized to 480×480, following the original paper [27]. We use the MiT-B3 variant of GeminiFusion [61]. Table 1 details our evaluation on DFormer and GeminiFusion.

On DFormer, our GeomPrompt module achieves an mIoU of 47.8 and a PA of 81.6, outperforming by +6.1 mIoU the RGB-only baseline, which yields an mIoU of 41.7. Furthermore, GeomPrompt performs competitively with state-of-the-art explicit monocular depth estimators, as it surpasses Depth Anything 2 (44.0 mIoU), Metric3Dv2

Table 1. Evaluation on DFormer and GeminiFusion. GT Depth is shown as an upper bound. Best non-upper-bound results for each benchmark are in bold.

	DFormer		GeminiFusion	
Method	**mIoU ↑**	**PA ↑**	**mIoU ↑**	**PA ↑**
GT Depth	51.2	83.4	52.7	82.8
RGB-only	41.7	78.3	43.4	78.8
DA2	44.0	80.8	**47.7**	**81.4**
DA2 [Hypersim]	47.5	**81.8**	44.5	79.6
Metric3Dv2	46.6	**81.8**	46.6	80.6
★ GeomPrompt	**47.8**	81.6	46.4	80.3

(46.6 mIoU), and the DA2 Hypersim checkpoint (47.5 mIoU).

When evaluated on the GeminiFusion backbone, GeomPrompt attains an mIoU of 46.4 and a PA of 80.3, providing an improvement of +3.0 mIoU over the RGB-only control baseline (43.4 mIoU). The learned geometric prompt performs comparably to the Metric3Dv2 estimator (46.6 mIoU) and comfortably outperforms the DA2 Hypersim checkpoint (44.5 mIoU), although the standard DA2 checkpoint achieves the highest zero-shot depth performance in this specific setup with an mIoU of 47.7.

This demonstrates that GeomPrompt successfully learns a task-relevant geometric representation strictly from segmentation supervision, and is competitive with explicitly trained monocular depth models across different multimodal fusion paradigms.

For context, we also compare to prior missing modality methods that train the segmentation model. On SUN RGB-D, M3L reports 41.31 mIoU for RGB inference with missing depth [36], and OS-MD reports 43.64 mIoU under incomplete modality evaluation [58]. GeomPrompt achieves 47.8 mIoU with DFormer and 46.4 mIoU with GeminiFusion under missing depth, despite training only a lightweight prompt module rather than the segmenter.

4.4. GeomPrompt-Recovery Under Depth Failures

To evaluate the robustness of GeomPrompt-Recovery (GPR), we subjected the model to a comprehensive suite of simulated depth degradations. The full corruption family includes quantize, hole, dropout, noise, blur, banding, and scale shifting. For this detailed analysis, we focus on three primary corruptions that commonly afflict real-world depth sensors [15, 22, 23].

The first degradation is quantization, which simulates limited sensor precision by mapping the depth values to discrete bins over the $[0, 255]$ range. The number of bins is bounded and inversely proportional to the severity so that higher severities yield significantly coarser depth levels [9, 40].

Table 2. Recovered gains across different degradation types.

Degradation	**Mean mIoU Gain ↑**	**High-Severity Gain ↑**
Quantization	+1.4	+2.3
Dropout	+2.0	+1.5
Noise	+2.5	+3.6

The second degradation is dropout, which mimics missing depth readings caused by reflective surfaces or sensor range limits by independently setting pixels to zero. The probability of a pixel being zeroed out is proportional to the severity. Unmasked pixels remain unchanged [10, 34, 73].

Our third degradation is noise, which simulates general sensor inaccuracy via additive zero-mean Gaussian noise. The noise is applied pixelwise with a standard deviation proportional to the severity, causing saturation at the 0 and 255 boundaries at higher severities [1, 22, 23].

Corruptions are applied globally across the depth maps. The clean depth is first converted to a uint8 format. The specific corruption is then executed in float32 precision, with the severity parameter strictly clamped to the $[0, 1]$ range. The final output is clipped and cast back to uint8.

Both the degraded depth and the GPR recovered prompt are passed into DFormer to evaluate segmentation results. Figure 3 illustrates the mIoU performance of the baseline degraded depth versus the recovered prompt across severities ranging from 0.1 to 0.9. Across all evaluated corruptions, GPR consistently acts as a buffer against catastrophic failure, maintaining higher mIoU than the raw degraded depth. Table 2 summarizes the mIoU improvements yielded by the recovered prompt compared to the degraded depth baseline.

The results indicate that GPR effectively recovers task-relevant geometric cues even when the input depth is severely compromised. Notably, the performance gap between the recovered prompt and the degraded depth generally widens as the severity increases. For example, while high-severity noise causes severe pixel saturation, GPR is still able to recover up to +3.6 mIoU.

This validates our core hypothesis that by predicting a bounded residual correction from the RGB input, GPR can partially restore the structured prior that is useful for the frozen segmenter. Because it is trained purely under a segmentation objective, GPR bypasses the need for explicit metric depth reconstruction, allowing it to dynamically compensate for structural failures in the depth map using contextual visual cues.

4.5. Analysis

4.5.1. Inference Cost

Since standard metric depth estimators are often computationally heavy [46, 69], we evaluate the efficiency of Geom-

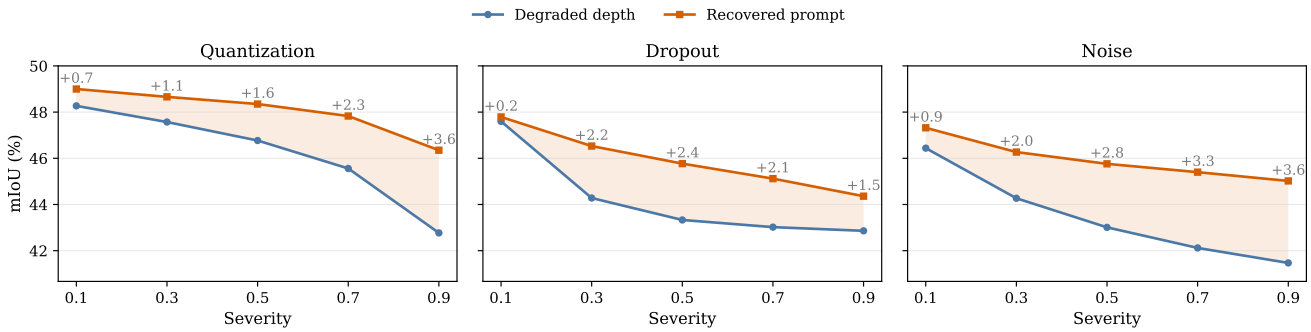


Figure 3. Degraded depth vs. our recovered prompt across severities and degradation types.

Table 3. Efficiency comparison across methods. Lower is better.

Model	Lat. (ms) ↓	FLOPs (G) ↓	Params (M) ↓
DA2	38.3	280.7	97.5
Metric3Dv2	71.9	315.2	37.5
★ GeomPrompt	**7.8**	**44.0**	**23.4**

Prompt against Depth Anything 2 (ViT-B) and Metric3Dv2 (ViT-S). The evaluation uses a batch size of 1 on a single A40 GPU with fp16 autocast precision. Latency includes only model execution time. Parameter counts reflect total inference-time parameters, and FLOPs are computed per forward pass on the respective preprocessed tensors. Table 3 details our efficiency comparison results.

GeomPrompt is substantially lighter than monocular depth baselines, as at 7.8 ms per frame, it is nearly 5 times faster than DA2 (38.3 ms) and over 9 times faster than Metric3Dv2 (71.9 ms). Furthermore, GeomPrompt requires only 44.0 GFLOPs and 23.4 million parameters compared to DA2's 280.7 GFLOPs and 97.5 million parameters. This lightweight profile suggests that GeomPrompt can be practical as an inline plugin for real-time embodied agents in environments with unreliable sensor depth.

4.5.2. Naive Depth Channel Controls

To verify GeomPrompt learns genuine geometric representations rather than exploiting low-level statistics, we evaluate it against four handcrafted pseudo-depth baselines. Luminance Gray (grayscale intensity) tests if brightness replaces geometry [3, 18]. Canny Distance Zero Blend blends a normalized Canny edge distance transform with zero depth to test edge-distance cues [5, 17]. Laplacian Edges uses an absolute Laplacian response normalized to $[0, 255]$ as a second-order structural prior [37]. Lastly, Scharr Distance Zero Blend applies an L2 distance transform to thresholded Scharr gradients, followed by normalization, Gaussian smoothing, and zero-depth blending for a smooth boundary prior [17, 45]. Table 4 details our results against these baselines.

Table 4. Naive baselines and references, separated by segmenter. Best results in each segmenter group are in bold.

Method	DFormer ↑	GeminiFusion ↑
RGB-only	41.7	43.4
Luminance Gray	25.4	43.2
Laplacian Edges	40.3	42.4
Scharr	39.8	43.0
Canny	38.6	43.5
★ GeomPrompt	**47.8**	**46.4**

Handcrafted pseudo-depth proxies consistently fail to surpass the RGB-only baseline across both backbones. On DFormer, naive controls degrade the 41.7 mIoU RGB baseline, dropping to 25.4 mIoU for luminance and peaking at 40.3 mIoU for Laplacian edges. Similarly, GeminiFusion's 43.4 mIoU baseline sees naive controls plateau between 42.4 and 43.5 mIoU. Conversely, GeomPrompt achieves 47.8 mIoU (DFormer) and 46.4 mIoU (GeminiFusion). The failure of classical structural cues confirms GeomPrompt extracts task-relevant geometric embeddings rather than merely highlighting edges or intensity gradients.

4.5.3. Training Ablations

We ablate key architectural and training components of GeomPrompt to isolate their contributions to the final performance on the SUN RGB-D test set. Figure 4 shows the mIoU of all ablations with respect to the baseline.

A constant residual scale ($s = 80$) is maintained, which isolates the value of the residual-scale curriculum by removing it and starting at maximum correction strength from epoch 1. The 1.4 mIoU drop suggests that linearly scaling the residual contribution during training acts as a stabilizer for prompt construction.

The PromptAdapter is disabled, which causes a modest but noticeable degradation, showing its utility in refining the predicted prompt to better match the expected distribution of the frozen segmenter.

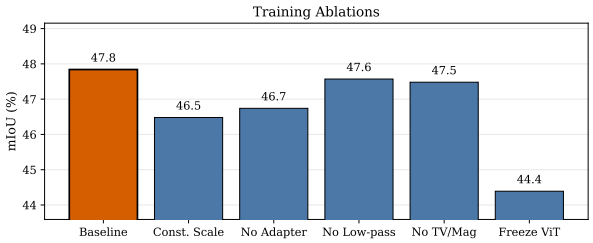


Figure 4. mIoU of training ablations vs. our baseline.

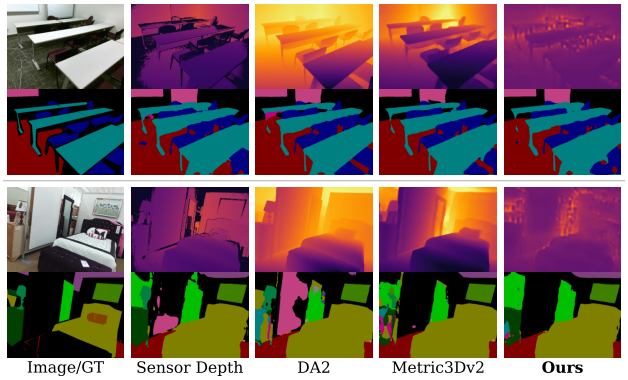


Figure 5. Qualitative comparison of segmentation outputs and geometric representations.

The low-pass projection is removed, which tests whether suppressing high-frequency residual structure is important for effective prompt construction. The resulting performance drops only slightly by 0.2 mIoU, indicating that the low-pass projection is not a primary driver of the gains.

The TV and magnitude regularization are disabled, which tests whether explicit smoothness and amplitude control are necessary for learning useful prompts. This causes a minor 0.3 mIoU degradation, suggesting that GeomPrompt does not rely heavily on them to achieve its improvement.

The ViT encoder is frozen, which tests whether end-to-end ViT feature adaptation is necessary. This results in a severe performance drop of 3.5 mIoU compared to the baseline, confirming that the base visual features must adapt to effectively synthesize geometric prompts.

4.6. Qualitative Results

We qualitatively evaluate our prompts to illustrate how they adapt to the absence or degradation of metric depth.

Figure 5 shows that GeomPrompt produces segmentation outputs competitive with depth-based alternatives while generating prompts that differ visibly from metric depth, emphasizing boundaries, planar structure, and semantic grouping. Figure 6 highlights this distinction by comparing our prompt and GT depth across images.

Figure 7 demonstrates the efficacy of GeomPrompt-Recovery (GPR) when subjected to severe sensor failures. GPR successfully processes the corrupted data to output a recovered representation that effectively smooths out high-frequency artifacts and appears to restore object boundaries.

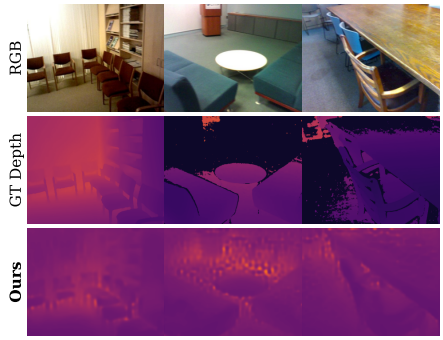


Figure 6. Visual comparison of our synthesized geometric prompt against the original RGB input and ground truth depth.

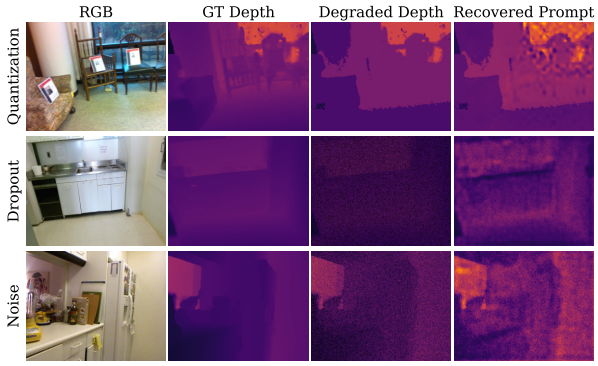


Figure 7. GeomPrompt-Recovery under simulated depth failures.

5. Conclusion

In this work, we presented GeomPrompt and GeomPrompt-Recovery as lightweight cross-modal adaptation modules for RGB-D perception under missing and degraded depth. Instead of reconstructing metric depth, the method learns task-driven geometric prompts from downstream supervision alone, enabling a frozen multimodal segmenter to remain effective when one sensing stream is unavailable or corrupted. GeomPrompt improves over RGB-only inference while remaining competitive with monocular depth estimators, and GeomPrompt-Recovery improves robustness under simulated sensor failures. These results suggest that task-driven cross-modal compensation can be a practical strategy for robust multimodal segmentation in embodied systems, especially when real-world deployment requires efficiency and graceful degradation under unreliable sensing. Future work could study whether this prompting view extends to other multimodal perception tasks useful for embodied AI, such as mapping, navigation, or manipulation.

References

[1] Fahira Afzal Maken, Sundaram Muthu, Chuong Nguyen, Changming Sun, Jinguang Tong, Shan Wang, Russell Tsuchida, David Howard, Simon Dunstall, and Lars Petersson. Improving 3D Reconstruction Through RGB-D Sensor Noise Modeling. *Sensors*, 25(3):950, 2025. 6

[2] Shashank Agnihotri, Julia Grabinski, and Margret Keuper. Improving Feature Stability During Upsampling – Spectral Artifacts and the Importance of Spatial Context. In *Computer Vision – ECCV 2024*, pages 357–376. Springer Nature Switzerland, Cham, 2025. 3

[3] Jonathan T. Barron and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–341, 2012. 7

[4] Minh Bui and Kostas Alexis. Diffusion-based RGB-D Semantic Segmentation with Deformable Attention Transformer. In *2025 IEEE International Conference on Advanced Robotics (ICAR)*, pages 404–411, San Juan, Argentina, 2025. IEEE. 3

[5] John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 7

[6] Adriano Cardace, Luca De Luigi, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Plugging Self-Supervised Monocular Depth into Unsupervised Domain Adaptation for Semantic Segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1999–2009, Waikoloa, HI, USA, 2022. IEEE. 3

[7] Siyu Chen, Ting Han, Changshe Zhang, Weiquan Liu, Jinhe Su, Zongyue Wang, and Guorong Cai. Depth Matters: Exploring Deep Interactions of RGB-D for Semantic Segmentation in Traffic Scenes. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7834–7841, Hangzhou, China, 2025. IEEE. 2, 3

[8] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning Semantic Segmentation From Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1841–1850, Long Beach, CA, USA, 2019. IEEE. 3

[9] Benjamin Choo, Michael Landau, Michael DeVore, and Peter Beling. Statistical Analysis-Based Error Models for the Microsoft KinectTM Depth Sensor. *Sensors*, 14(9):17430–17450, 2014. 6

[10] Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning Depth Estimation for Transparent and Mirror Surfaces. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9210–9221, Paris, France, 2023. IEEE. 6

[11] Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. Semantic Information for Robot Navigation: A Survey. *Applied Sciences*, 10(2):497, 2020. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, 2009. IEEE. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020. 3

[14] Siqi Du, Weixi Wang, Renzhong Guo, Ruisheng Wang, and Shengjun Tang. AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7608–7615, Seattle, WA, USA, 2024. IEEE. 2

[15] Alexandre Duarte, Francisco Fernandes, João M. Pereira, Catarina Moreira, Jacinto C. Nascimento, and Joaquim Jorge. Selfredepth: Self-supervised real-time depth restoration for consumer-grade sensors. *Journal of Real-Time Image Processing*, 21(4):124, 2024. 6

[16] Tiyu Fang, Zhen Liang, Xiuli Shao, Zihao Dong, and Jinping Li. Depth Removal Distillation for RGB-D Semantic Segmentation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2405–2409, Singapore, Singapore, 2022. IEEE. 3

[17] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 8 (1):415–428, 2012. 7

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 7

[19] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3827–3837, Seoul, Korea (South), 2019. IEEE. 2, 3

[20] Zhangxuan Gu, Li Niu, Haohua Zhao, and Liqing Zhang. Hard Pixel Mining for Depth Privileged Semantic Segmentation. *IEEE Transactions on Multimedia*, 23:3738–3751, 2021. 2, 3

[21] Yanrong Guo and Tao Chen. Semantic segmentation of RGBD images based on deep depth regression. *Pattern Recognition Letters*, 109:55–64, 2018. 3

[22] Azmi Haider and Hagit Hel-Or. What Can We Learn from Depth Camera Sensor Noise? *Sensors*, 22(14):5448, 2022. 2, 6

[23] Ying He, Bin Liang, Yu Zou, Jin He, and Jun Yang. Depth Errors Analysis and Correction for Time-of-Flight (ToF) Cameras. *Sensors*, 17(1):92, 2017. 2, 6

[24] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving Semi-Supervised and Domain-Adaptive Semantic Segmentation with Self-Supervised Depth Estimation. *International Journal of Computer Vision*, 131(8):2070–2096, 2023. 3

[25] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 2, 3, 5

[26] Xuanlun Huang, Chenyang Wu, Xiaolan Xu, Baishun Wang, Sui Zhang, Chihchiang Shen, Chiennan Yu, Jiaxing Wang, Nan Chi, Shaohua Yu, and Connie J. Chang-Hasnain. Polarization structured light 3D depth image sensor for scenes with reflective surfaces. *Nature Communications*, 14(1): 6855, 2023. 2

[27] Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. GeminiFusion: Efficient Pixel-wise Multimodal Fusion for Vision Transformer, 2024. 2, 5

[28] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson Lau, and Thomas S. Huang. Geometry-Aware Distillation for Indoor Semantic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2864–2873, Long Beach, CA, USA, 2019. IEEE. 2, 3

[29] Faisal Khan, Saqib Salahuddin, and Hossein Javidnia. Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review. *Sensors*, 20(8):2272, 2020. 2

[30] Juan Lagos and Esa Rahtu. SemSegDepth: A Combined Model for Semantic Segmentation and Depth Completion. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 155–165, Online Streaming, — Select a Country —, 2022. SCITEPRESS - Science and Technology Publications. 3

[31] Dong Lao, Fengyu Yang, Daniel Wang, Hyoungseob Park, Samuel Lu, Alex Wong, and Stefano Soatto. On the Viability of Monocular Depth Pre-training for Semantic Segmentation. In *Computer Vision – ECCV 2024*, pages 340–357. Springer Nature Switzerland, Cham, 2025. 3

[32] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: Privileged Adversarial Learning from Simulation, 2018. 3

[33] Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking Multi-Modal Semantic Segmentation Under Sensor Failures: Missing and Noisy Modality Robustness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1567–1577, Nashville, TN, USA, 2025. IEEE. 2, 3

[34] Boqian Liu, Haojie Li, Zhihui Wang, and Tianfan Xue. Transparent Depth Completion Using Segmentation Features. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(12):1–19, 2024. 6

[35] Yiming Lu, Bin Ge, Chenxing Xia, Xu Zhu, Mengge Zhang, Mengya Gao, Ningjie Chen, Jianjun Hu, and Junjie Zhi. FCEGNet: Feature calibration and edge-guided MLP decoder Network for RGB-D semantic segmentation. *Computer Vision and Image Understanding*, 260:104448, 2025. 2

[36] Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing Modality Robustness in Semi-Supervised Multi-Modal Semantic Segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1009–1019, Waikoloa, HI, USA, 2024. IEEE. 3, 6

[37] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 7

[38] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and Checkerboard Artifacts. *Distill*, 1(10): 10.23915/distill.00003, 2016. 3

[39] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal Monocular Metric Depth Estimation Made Simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026. 3

[40] Ishraq Rached, Rafika Hajji, Tania Landes, and Rashid Haffadi. StructScan3D v1: A First RGB-D Dataset for Indoor Building Elements Segmentation and BIM Modeling. *Sensors*, 25(11):3461, 2025. 6

[41] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the Spectral Bias of Neural Networks. 2018. 3

[42] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168, Montreal, QC, Canada, 2021. IEEE. 3

[43] Sonia Raychaudhuri and Angel X. Chang. Semantic Mapping in Indoor Embodied AI – A Survey on Advances, Challenges, and Future Directions, 2025. 2

[44] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10892–10902, Montreal, QC, Canada, 2021. IEEE. 5

[45] Hanno Scharr. Optimal Filters for Extended Optical Flow. In *Complex Motion*, pages 14–29. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. 7

[46] Claudio Schiavella, Lorenzo Cirillo, Lorenzo Papa, Paolo Russo, and Irene Amerini. Efficient attention vision transformers for monocular depth estimation on resource-limited hardware. *Scientific Reports*, 15(1):24001, 2025. 6

[47] Daniel Seichter, Sohnke Benedikt Fischedick, Mona Kohler, and Horst-Michael Grob. Efficient Multi-Task RGB-D Scene Analysis for Indoor Environments. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, Padua, Italy, 2022. IEEE. 2

[48] Leon Sick, Dominik Engel, Pedro Hermosilla, and Timo Ropinski. Unsupervised Semantic Segmentation Through Depth-Guided Feature Correlation and Sampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3637–3646, Seattle, WA, USA, 2024. IEEE. 3

[49] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark

suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, Boston, MA, USA, 2015. IEEE. 5

[50] Martin Stommel, Michael Beetz, and Weiliang Xu. Inpainting of Missing Values in the Kinect Sensor's Depth Maps Based on Background Estimates. *IEEE Sensors Journal*, 14 (4):1107–1116, 2014. 2

[51] Jingwen Sun, Jing Wu, Ze Ji, and Yu-Kun Lai. A Survey of Object Goal Navigation. *IEEE Transactions on Automation Science and Engineering*, 22:2292–2308, 2025. 2

[52] Wenbo Sun, Zhi Gao, Jinqiang Cui, Bharath Ramesh, Bin Zhang, and Ziyao Li. Semantic Segmentation Leveraging Simultaneous Depth Estimation. *Sensors*, 21(3):690, 2021. 3

[53] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. DADA: Depth-Aware Domain Adaptation in Semantic Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7363–7372, Seoul, Korea (South), 2019. IEEE. 3

[54] Zifu Wan, Pingping Zhang, Yuhao Wang, Silong Yong, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1734–1744, Tucson, AZ, USA, 2025. IEEE. 2

[55] Changshuo Wang, Chen Wang, Weijun Li, and Haining Wang. A brief survey on RGB-D semantic segmentation using deep learning. *Displays*, 70:102080, 2021. 2

[56] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep Multimodal Fusion by Channel Exchanging, 2020. 2

[57] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal Token Fusion for Vision Transformers, 2022. 2

[58] Shicai Wei, Yang Luo, and Chunbo Luo. One-stage Modality Distillation for Incomplete Multimodal Learning, 2023. 6

[59] Suhan Woo, Junhyuk Hyun, Suhyeon Lee, and Euntai Kim. Real-time RGB-D Semantic Segmentation With Scale-invariant Depth Encoding and Noise-robust Fusion. *International Journal of Control, Automation and Systems*, 23(12): 3649–3661, 2025. 2, 3

[60] Zongwei Wu, Zhuyun Zhou, Guillaume Allibert, Christophe Stolz, Cédric Demonceaux, and Chao Ma. Transformer fusion for indoor RGB-D semantic segmentation. *Computer Vision and Image Understanding*, 249:104174, 2024. 2

[61] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, 2021. 5

[62] Xinhua Xu, Jinfu Liu, and Hong Liu. Interactive Efficient Multi-Task Network for RGB-D Semantic Segmentation. *Electronics*, 12(18):3943, 2023. 2

[63] Xinhua Xu, Hong Liu, Jianbing Wu, and Jinfu Liu. PDDM: Pseudo Depth Diffusion Model for RGB-PD Semantic Segmentation Based in Complex Indoor Scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):8969–8977, 2025. 3

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2, 2024. 2, 3, 5

[65] Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. DFormer: Rethinking RGBD Representation Learning for Semantic Segmentation, 2023. 2, 5

[66] Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. DFormerv2: Geometry Self-Attention for RGBD Semantic Segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19345–19355, Nashville, TN, USA, 2025. IEEE. 2

[67] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24 (12):14679–14694, 2023. 2

[68] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering Arbitrary-Modal Semantic Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1136–1147, Vancouver, BC, Canada, 2023. IEEE. 3

[69] Jiuling Zhang, Yurong Wu, and Huilong Jiang. Survey on monocular metric depth estimation. *Computers*, 14(11), 2025. 6

[70] Richard Zhang. Making convolutional networks shift-invariant again. *CoRR*, abs/1904.11486, 2019. 4

[71] Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the Value of Every Modality: Towards Efficient and Resilient Modality-Agnostic Semantic Segmentation. In *Computer Vision – ECCV 2024*, pages 192–212. Springer Nature Switzerland, Cham, 2025. 3

[72] Li Zhong, Chi Guo, Jiao Zhan, and JingYi Deng. Attention-based fusion network for RGB-D semantic segmentation. *Neurocomputing*, 608:128371, 2024. 2, 3

[73] Luyang Zhu, Arsalan Mousavian, Yu Xiang, Hammad Mazhar, Jozef van Eenbergen, Shoubhik Debnath, and Dieter Fox. RGB-D Local Implicit Function for Depth Completion of Transparent Objects, 2021. arXiv:2104.00622. 6